Clarify or Answer: Reinforcement Learning for Agentic VQA with Context Under-specification

Zongwan Cao[1]* **Bingbing Wen**[1]* **Lucy Lu Wang**[1,2]
[1]University of Washington
[2]Allen Institute for AI
{zongwanc, bingbw, lucylw}@uw.edu

Abstract

Real-world visual question answering (VQA) may be *context-dependent*: an image-question pair could be under-specified, such that the correct answer depends on external information that is not observable in the image. In such cases, directly answering can lead to confident but incorrect predictions. We propose Clarify-or-Answer (CoA), an ask-or-answer agent that separately models the decision to ask or answer, and what to ask if needed. CoA first determines whether clarification is necessary; if so, it generates one or more clarification questions to iteratively enrich the context until ambiguity is resolved. To validate this framework, we introduce CONTEXTCLARIFY, a dataset of ambiguous VQA questions along with a contrast set of non-ambiguous questions. We further introduce GRPO-CR (Clarification Reasoning), a reinforcement learning approach that optimizes clarification question generation with multiple reward signals encouraging well-formed, focused, non-trivial questions that resolve ambiguity. Across three VLMs and three datasets, CoA achieves consistent improvements at both the module and system levels, with a single clarification turn providing substantial gains and recursive multi-turn execution offering modest additional improvements in select settings.[1]

1 Introduction

Visual Question Answering (VQA) asks a model to answer a natural language question given an image (Agrawal et al., 2016). Real-world VQA is an interactive decision problem: an agent must decide whether it can answer reliably from the image alone, or whether it should first *seek missing context* that is not visually observable (Luo et al., 2024). When the input is underspecified, directly answering tends to produce confident but incorrect predictions. For example, in Figure 1, the question *When was this photo taken, the first or second half of the year?* cannot be answered without knowing the *location* of the photographer. A reliable VQA agent should therefore adapt its behavior: it should answer immediately when the VQA is sufficiently specified, and otherwise ask a targeted clarification question to obtain the missing information before answering.

Recent research has examined multiple sources of ambiguity in VQA, including perceptual ambiguity and visual illusions (Guan et al., 2023), referential and focus-based uncertainty (Chen et al., 2025), multilingual and context-dependent ambiguity (Luo et al., 2024; Wang et al., 2025; Wen et al., 2024), and robustness to rephrasings or visual corruption (Shah et al., 2019; Farhan et al., 2024). Interactive formulations, e.g., ClearVQA, encourage asking clarification questions before answering when inputs are ambiguous (Jian et al., 2025). However, in prior work, ambiguity typically resides in the question phrasing or image content itself, not in the combination of both. This distinction inspires our focus on a complementary setting: underspecified VQA pairs, where the image and question individually are clear, but together yield multiple potentially valid answers depending on external context.

*Equal contribution
[1]Our dataset and code can be found at https://github.com/zwverse/clarify-or-answer

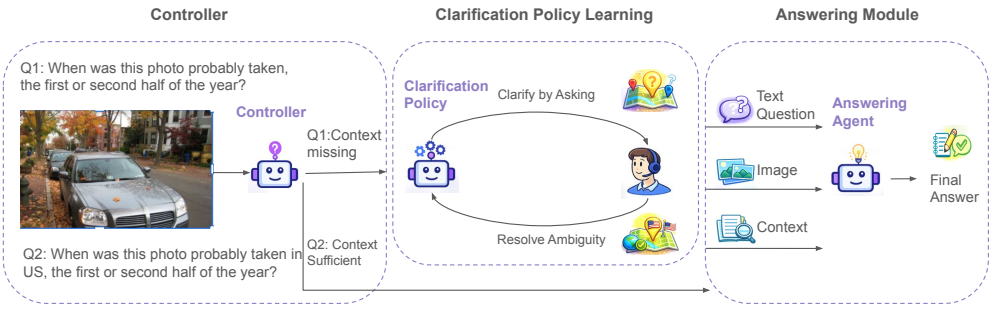


Figure 1: System overview of Clarify-or-Answer (CoA). The controller decides whether a visual question requires clarification. Ambiguous queries with missing context trigger a clarification step (clarification policy learning), while context-sufficient queries are directly forwarded to the VQA answering module, which produces the final answer.

In response, we propose CoA (Clarify-or-Answer), an agentic framework that explicitly separates the decision of whether to answer from clarification question generation. CoA consists of three components: (i) a *controller* that decides whether the agent should answer or ask for clarification, (ii) a *clarification policy* that generates a single targeted clarification question when needed, and (iii) an *answering model* that produces the final answer using the available context (Figure 1).

We further study how to train the clarification policy of CoA to ask *effective* questions rather than generic restatements. We curate CONTEXTCLARIFY, a dataset of 266 ambiguous image-question pairs derived from CODIS (Luo et al., 2024) and supplemented with manually sourced examples, each annotated with a human-verified, open-ended clarification question. To train and evaluate the controller, we construct a contrast set by augmenting each ambiguous instance with context-completed and irrelevant-context variants, yielding cases where clarification is necessary and where it is not. For learning to ask clarification questions, we introduce GRPO-CR, a reinforcement learning approach based on GRPO (Shao et al., 2024) that optimizes a multi-signal reward capturing question format, relevance to missing context, novelty, similarity to human clarifications, and ambiguity resolution potential.

We evaluate CoA at both the module level (controller decisions; clarification quality) and the system level (overall VQA accuracy), comparing against prompting and supervised baselines. While CoA can operate as a single-step interaction, it also supports a recursive multi-turn paradigm where the agent iteratively asks one or more clarification questions to enrich the context until the ambiguity is resolved. Our results demonstrate that CoA outperforms prompting and supervised finetuning baselines; a single turn provides substantial VQA accuracy gains on ambiguous questions, and multi-turn execution offers small additional boosts to accuracy in most settings.

In summary, our contributions are:

- We propose CoA, an ask-or-answer agent that separates *when to ask* (controller) from *what to ask* (clarification policy), enabling adaptive decision making between clarifying and answering.
- We define the task of VQA clarification question generation in cases of context underspecification. We introduce CONTEXTCLARIFY, a dataset of 266 ambiguous VQA instances with human-verified open-ended clarifications and an additional 266 non-ambiguous contrast instances for training and evaluating ask-or-answer decisions.
- We develop GRPO-CR, a GRPO-based reinforcement learning method that improves clarification quality. CoA with GRPO-CR consistently outperforms prompting and finetuned baselines in controller logic, clarification policy execution, and final VQA accuracy across all tested datasets.

2 Related Work

Visual Ambiguity Visual ambiguity may arise from incomplete visual cues or distracting noise in a scene (Denison et al., 2018). Much recent work on ambiguity detection focuses on benchmark creation and assessing how well off-the-shelf vision-language models (VLMs) handle different types of ambiguity (Yue et al., 2024; Yao et al., 2025; Yang et al., 2025). Some studies examine ambiguities caused by optical illusions (Cui et al., 2023; Fu et al., 2023;

Guan et al., 2023). Datasets like CODIS (Luo et al., 2024) evaluate whether models interpret images correctly across different contextual settings, while MUCAR (Wang et al., 2025) introduces dual-ambiguity scenarios spanning both textual and visual dimensions. Audits of VQA benchmarks further demonstrate how underspecified temporal, spatial, or attribute scope introduces recurring visual ambiguity Qian et al. (2026). Recently, ClearVQA (Jian et al., 2025) proposes an interactive dataset where models must ask clarification questions when faced with ambiguous visual questions, and introduces a DPO-based training method to promote clarification-seeking behavior.

Our work is complementary to and differs from ClearVQA (Jian et al., 2025) in several ways. In our setting, neither the image nor text is inherently ambiguous; instead, the VQA pair is underspecified, allowing multiple reasonable answers depending on external context. Thus, while both tasks involve clarification question generation, the underlying use cases are distinct. Further, ClearVQA generates Yes/No clarification questions, whereas our work focuses on open-ended clarification. Methodologically, we adopt GRPO reinforcement learning, in contrast to the DPO framework (Rafailov et al., 2023) used in ClearVQA.

Ambiguity Resolution through Question Generation Beyond benchmark construction, a growing body of research explores how models proactively ask clarification questions to resolve uncertainty. For text-based tasks, clarification question generation has been studied in dialogue and QA settings, where models are trained to request missing information rather than produce incomplete or incorrect answers (Rao & Daumé, 2018; Aliannejadi et al., 2019; Wen et al., 2025). Recent language model work has extended this idea, using preference optimization or reinforcement learning to encourage clarification seeking in conversational agents (Zhang et al., 2024; Li et al., 2025; Zhang & Choi, 2023; Wen et al., 2023).

However, work in multimodal domains remains limited. While ClearVQA (Jian et al., 2025) pioneers clarification question generation in VQA and recent approaches explore candidate answer generation under uncertainty (Wang & Itou, 2026), our work extends this prior work on ambiguity resolution to context-dependent and ambiguous VQA pairs, providing a dataset and training method to enhance performance on this task.

3 Methodology

3.1 Problem Setup

We study *context-dependent* VQA, where answering a question Q_0 given an image V may require external context c that is not visually observable. Following prior work on contextual ambiguity (Luo et al., 2024), we consider five common underspecification types: (1) location and orientation, (2) temporal information, (3) cultural background, (4) attributes, and (5) relationships. Formally, the agent must decide whether the available context $\mathcal{X}_t$ (where $\mathcal{X}_0 = \{Q_0, V\}$) is sufficient to produce a correct answer A, or whether it must first elicit missing information. We model this as a decision problem where the agent selects an action $u \in \{\textsc{Answer}, \textsc{Clarify}\}$. If the agent chooses $\textsc{Clarify}$, it generates a clarification question C_t to retrieve a response R_t, thereby augmenting the context $\mathcal{X}_{t+1} = \{\mathcal{X}_t, C_t, R_t\}$.

3.2 CoA: Clarify-or-Answer Agent

We propose CoA to resolve contextual ambiguity in this setting. CoA consists of three components:

- **Controller** $\pi_{\text{ctrl}}(u \mid \mathcal{X}_t)$: Selects the agent action $u \in \{\textsc{Answer}, \textsc{Clarify}\}$
- **Clarification policy** $\pi_{\text{clr}}(C_t \mid \mathcal{X}_t)$: Generates a targeted clarification question C_t when $u = \textsc{Clarify}$
- **Answering model** $\pi_{\text{ans}}(A \mid \mathcal{X}_t)$: Outputs the final answer A conditioned on the accumulated context $\mathcal{X}_t$

At inference time, CoA evaluates the current context $\mathcal{X}_t$ using the controller. If $u = \textsc{Answer}$, the agent provides the final answer A. If $u = \textsc{Clarify}$, it generates C_t, receives R_t, and updates the context to $\mathcal{X}_{t+1}$. Decoupling the controller and policy avoids conflicting supervision: the controller requires training on mixed contrast sets to judge ambiguity, while the policy is optimized via RL strictly on missing-context instances.

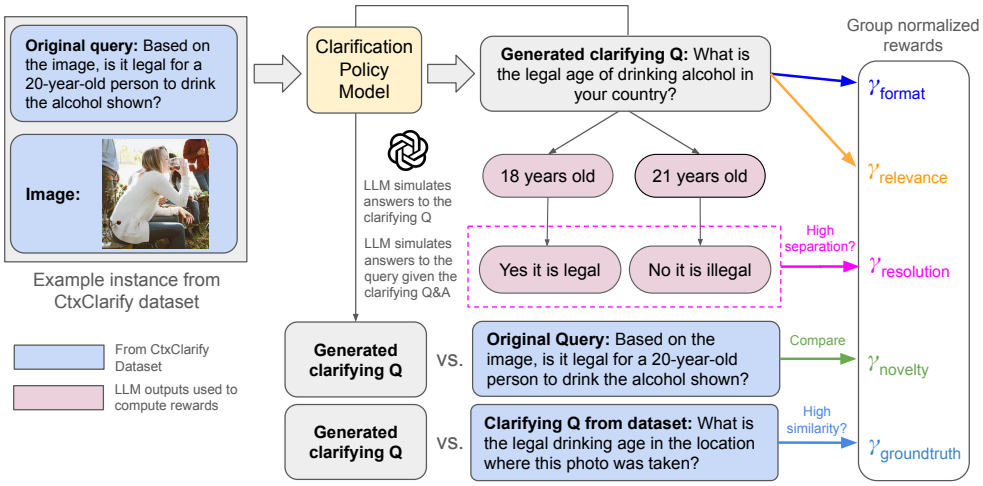


Figure 2: Training pipeline for our method GRPO-CR. The clarification policy model generates clarification questions from image-query pairs, which are evaluated by multi-dimensional rewards and updated via GRPO.

3.3 Controller Learning Objective

The controller is trained to predict whether missing context is required to answer reliably. We formulate this as binary classification over $u \in \{\text{ANSWER}, \text{CLARIFY}\}$ and optimize the standard cross-entropy loss $\mathcal{L}_{\text{ctrl}} = -\log \pi_{\text{ctrl}}(u^* \mid \mathcal{X}_0)$, where u^* is the ground-truth ask-or-answer label for the initial state.

3.4 Clarification Policy Learning

For ambiguous inputs, the clarification policy is trained to generate a question C_t that effectively elicits the missing contextual factor c.

Reinforcement learning with GRPO-CR. We introduce GRPO-CR (Clarification Reasoning) (see Figure 2), a reinforcement learning approach built on Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each input $\mathcal{X}_t$, the policy samples a group of N candidate clarifications $\{C_t^{(i)}\}_{i=1}^N$ and receives a scalar reward $R(C_t^{(i)})$. We compute within-group relative advantages as:

$$\hat{A}_i = R(C_t^{(i)}) - \frac{1}{N}\sum_{j=1}^{N} R(C_t^{(j)})$$

and optimize the policy loss:

$$\mathcal{L}_{\text{GRPO-CR}} = -\mathbb{E}_i \left[\log \pi_{\text{clr}} \left(C_t^{(i)} \mid \mathcal{X}_t \right) \cdot \hat{A}_i \right]$$

Reward Components Our reward design builds on prior work showing that effective clarification requires balancing structural validity and informativeness. For example, ALFA demonstrates the value of rewards capturing both format compliance and utility (Li et al., 2025), while GRIT highlights the need for explicit rewards in multimodal reasoning (Fan et al., 2025). In contrast to text-only or general reasoning settings, we introduce a tailored set of multimodal rewards designed for context underspecification in VQA.

The total reward $R(C_t)$ is a weighted sum of individual rewards $R(C_t) = \sum_k \lambda_k R_k(C_t)$ with $\lambda_k = 1$. Individual reward components are determined based on relative importance and a parameter sweep over reward coefficients (Appendix A).

- **Format reward (γ_{format})**: The format reward encourages outputs that are well-formed, concise clarification questions. Assigns $+0.5$ if C_t ends with a question mark and is under 50 words; otherwise -0.5.
- **Focused relevance reward ($\gamma_{\text{relevance}}$)**: This reward encourages clarification questions that explicitly target the missing context. We define a series of lexical patterns (e.g., *what*, *which*) and keywords associated with different categories of questions (e.g., *country*, *city* for cultural ambiguity) (full descriptions and lists in Appendix B). We assign higher

scores to questions that (i) follow clarification-oriented patterns and (ii) include any keywords associated with the type of ambiguity. C_t satisfying both conditions receive $+0.3$, those satisfying only one receive $+0.1$, and others receive -0.1.

- **Ambiguity resolution reward ($\gamma_{\text{resolution}}$)**: Assigns $+0.5$ if the response R_t leads to a different final answer A, indicating the ambiguity was resolved, and -0.3 otherwise.
- **Novelty reward (R_{novelty})**: This reward penalizes trivial rephrasings of the original question Q_0 by encouraging semantic diversity. We compute the Jaccard similarity J between the clarification question C_t and Q_0. If $J > 0.8$, the reward is -0.3; if $0.8 \geq J \geq 0.6$, the reward is -0.1; if $J < 0.6$, the reward is $+0.3$.
- **Ground-truth check reward ($\gamma_{\text{groundtruth}}$)**: This reward measures alignment between the generated clarification question and human-authored reference clarifications. It uses GPT-4o (OpenAI, 2024) to assign a similarity score in $[0, 1]$ between C_t and a human-authored reference.

We further validate our reward design via reward ablation studies to evaluate robustness. Details are reported in the Appendix C.

3.5 End-to-End CoA Execution

The CoA framework is evaluated under two execution settings:

Single-turn: The model performs a one-step decision on $\mathcal{X}_0$. It either generates a direct answer A or issues exactly one clarification question C_1. Upon receiving R_1, the framework produces the final answer A based on the augmented context $\mathcal{X}_1 = \{Q_0, V, C_1, R_1\}$.

Multi-turn: The clarification process is treated as a recursive loop. At each turn t, the model evaluates $\mathcal{X}_t$. If it selects CLARIFY, it generates C_t and receives R_t, forming $\mathcal{X}_{t+1} = \{\mathcal{X}_t, C_t, R_t\}$. This context is fed back into the CoA framework to re-evaluate the evidence. If the model further selects CLARIFY, new C_t and R_t are appended to the context. This continues until the model selects ANSWER or reaches a maximum turn limit T. If the turn limit T is reached, the current augmented context is used to produce a final answer.

4 CONTEXTCLARIFY Dataset

To validate CoA and GRPO-CR on underspecified VQA pairs, we introduce CONTEXTCLARIFY, a curated dataset of context-dependent VQA pairs annotated with open-ended clarification questions, and a corresponding contrast set with non-ambiguous VQA pairs.

Source Dataset Our dataset is derived from CODIS (Luo et al., 2024), which contains 222 images paired with 395 questions. CODIS questions exhibit diverse forms of contextual ambiguity, including categories such as *location and orientation*, *temporal information*, *cultural background*, *attributes*, and *relationships*. While this makes CODIS well suited for studying ambiguity, not all question-image pairs are appropriate for learning targeted clarification behavior, as some questions are vague or require resolving multiple missing factors.

Dataset Curation and Clarification Annotation From the original CODIS dataset, we curate a subset of 266 high-quality VQA items that each depend on a single missing contextual factor. Approximately 30% of CODIS pairs are discarded due to poor visual grounding or compound ambiguities. To improve clarity, we further replace about 10% CODIS images with carefully matched ambiguous images collected from the Internet. These replacements preserve the original contextual dependencies while providing stronger visual grounding.

For each retained item, we annotate one clarification question that explicitly targets missing context required to answer the question correctly. We first use GPT-4o to generate candidate clarification questions given the image, question, and known ambiguous contexts, and then manually review the outputs for acceptability. During review, annotators either accept the generated question without change or refine it to improve precision. This curated subset of CODIS along with our annotated clarification questions are included in CONTEXTCLARIFY.

Contrast Dataset Construction For each ambiguous item from before, we construct two variants: (i) a fully specified version where the missing relevant context is added into the question, yielding a question that no longer requires clarification, and (ii) a version

augmented with additional but irrelevant contextual information that does not resolve the ambiguity and still requires clarification. This procedure results in a mixture of questions that either do or do not require clarification that we use to train our Controller. Additional details on ambiguity categories, mixed-dataset construction rules, and illustrative examples are provided in Appendix D.

Final Dataset Statistics CONTEXTCLARIFY and its contrast set contain 532 VQA instances split 50/50 between ambiguous and non-ambiguous contrast items across 70% train, 15% validation, and 15% test splits. It spans five contextual ambiguity categories: *location and orientation*, *temporal information*, *cultural background*, *attributes*, and *relationships*. See Appendix D.3 for full breakdown.

5 Experimental Setup

5.1 Datasets, Models & Baselines

Datasets We use CONTEXTCLARIFY (abbreviated CtxClarify) for training and evaluation of CoA. To assess generalization beyond CONTEXTCLARIFY, we also evaluate on VizWiz (Chen et al., 2025) and ClearVQA (Jian et al., 2025), using their original train/dev/test splits. Both VizWiz and ClearVQA contain different types of visual ambiguity compared to CONTEXTCLARIFY.

Models We adopt three VLMs of varying scales: InternVL3-2B-Instruct (Zhu et al., 2025), Qwen2.5-VL-7B-Instruct (Qwen et al., 2024) and Qwen2.5-VL-3B-Instruct (Qwen et al., 2024) as backbones for the controller, clarification policy, and answering models in our framework (referred to as IVL-2B, Qw-3B, and Qw-7B, respectively). Additionally, to provide reference points against state-of-the-art commercial systems, we evaluate closed-weight frontier models GPT-4o (OpenAI et al., 2024) and Gemini-2.5-Flash (Comanici et al., 2025) under a zero-shot prompting configuration.

Baselines We compare CoA against baselines at both the system and module level. For the system-level baseline, we use (i) *prompting*, where the backbone model directly decides whether to answer or ask a clarification question, (ii) *always ask (AA_{GRPO})*, where the GRPO-trained clarification model bypasses the controller and always produces a clarification question, (iii) *in-context learning (ICL)* using 4 few-shot examples (Appendix E.3), and (iv) an *oracle setting* ($\text{CoA}_{\text{Oracle}}$), where the controller is assumed to be 100% accurate at providing ground-truth ask-or-answer labels, isolating downstream clarification policy performance.

For module-level baselines: (i) Controller module: we *prompt* the backbone model to directly classify whether a question is ambiguous or not. (ii) Clarification module: we compare against *prompting*, where the backbone model is directly prompted to generate a clarification question, and *supervised finetuning (SFT)*, where the backbone model is trained on CONTEXTCLARIFY to generate clarification questions.

5.2 Implementation Details

Controller training. We train controllers on CONTEXTCLARIFY (over the train split of ambiguous questions and the contrast set) with cross-entropy over {ANSWER, CLARIFY}. We implement supervised finetuning using MS-SWIFT (Zhao et al., 2024). Prompting baselines use a fixed decision prompt (Appendix I).

Clarification policy training. We train only on the train split of the ambiguous subset of CONTEXTCLARIFY. The policy described in §3.4 is optimized using GRPO to refine clarification strategies through relative reward signals.

End-to-end CoA execution. In the end-to-end setting, CoA either answers directly when the controller selects ANSWER or generates a question when the controller selects CLARIFY. When a clarification question is generated, we use GPT-4o to simulate a plausible user response to the question conditioned on the image (following prior work such as Jian et al. (2025)), and condition the answering model on all of the above when producing the final answer. This avoids reliance on oracle context while allowing us to assess whether the learned clarification strategies translate to improvements in downstream VQA accuracy.

All modules use the same backbone model and follow identical training and inference protocols to ensure fair comparisons. Further implementation details are in Appendix E.

Model	Attributes	Cultural Backgr.	Location & Orient.	Relation-ships	Temporal Info.	Overall
Closed-weight models (zero-shot)						
GPT-4o	0.611±0.138	0.690±0.102	0.630±0.080	0.714±0.177	0.483±0.072	0.619±0.030
Gemini-2.5-flash	0.630±0.080	0.571±0.177	0.556±0.138	0.690±0.102	0.383±0.072	0.560±0.030
Open-weight models (baselines and finetuned on CONTEXTCLARIFY *using CoA framework)*						
InternVL3-2B (Zero-shot)	0.093±0.080	0.048±0.102	0.056±0.138	0.143±0.177	0.033±0.072	0.087±0.017
InternVL3-2B (AA_{GRPO})	0.278±0.138	0.119±0.102	**0.407±0.080**	0.143±0.177	0.217±0.072	0.238±0.030
InternVL3-2B (ICL)	0.130±0.080	0.143±0.177	0.167±0.138	0.095±0.102	0.100±0.124	0.123±0.017
InternVL3-2B (CoA_{SFT})	0.185±0.080	0.119±0.102	0.130±0.080	0.190±0.102	0.083±0.072	0.127±0.017
InternVL3-2B ($CoA_{SFT+GRPO}$)	**0.389±0.138**	**0.310±0.102**	0.204±0.080	**0.286±0.177**	**0.367±0.072**	**0.313±0.017**
InternVL3-2B (CoA_{Oracle})	0.389±0.138	0.333±0.102	0.296±0.080	0.357±0.177	0.400±0.124	0.361±0.017
Qwen2.5-3B (Zero-shot)	0.315±0.080	0.095±0.102	0.278±0.138	**0.381±0.102**	0.100±0.124	0.242±0.017
Qwen2.5-3B (AA_{GRPO})	0.278±0.138	0.286±0.177	0.296±0.080	0.262±0.102	**0.350±0.124**	0.321±0.030
Qwen2.5-3B (ICL)	0.352±0.080	0.143±0.177	0.278±0.138	**0.381±0.102**	0.233±0.072	0.274±0.030
Qwen2.5-3B (CoA_{SFT})	**0.426±0.080**	0.310±0.102	0.352±0.080	0.143±0.177	**0.350±0.124**	0.321±0.030
Qwen2.5-3B ($CoA_{SFT+GRPO}$)	0.389±0.138	**0.381±0.102**	**0.463±0.080**	0.310±0.102	**0.350±0.124**	**0.381±0.030**
Qwen2.5-3B (CoA_{Oracle})	0.444±0.138	0.405±0.102	0.500±0.138	0.310±0.102	0.400±0.124	0.421±0.017
Qwen2.5-7B (Zero-shot)	0.259±0.080	0.452±0.102	0.315±0.080	0.262±0.102	0.083±0.072	0.266±0.017
Qwen2.5-7B (AA_{GRPO})	0.333±0.138	0.405±0.102	**0.389±0.138**	0.286±0.177	0.367±0.072	0.357±0.030
Qwen2.5-7B (ICL)	0.148±0.080	0.357±0.177	0.278±0.138	0.286±0.177	0.083±0.072	0.230±0.017
Qwen2.5-7B (CoA_{SFT})	0.296±0.080	0.524±0.102	0.352±0.080	0.310±0.102	0.250±0.124	0.317±0.017
Qwen2.5-7B ($CoA_{SFT+GRPO}$)	**0.389±0.138**	**0.571±0.177**	**0.389±0.138**	**0.333±0.102**	**0.400±0.124**	**0.409±0.017**
Qwen2.5-7B (CoA_{Oracle})	0.444±0.138	0.571±0.177	0.444±0.138	0.429±0.177	0.417±0.072	0.452±0.030

Table 1: VQA accuracy on CONTEXTCLARIFY. Our CoA-tuned models improve over their base counterparts across all question categories, in some cases approaching performance of frontier VLMs.

5.3 Evaluation Metrics

VQA accuracy We measure final answer correctness by reporting standard VQA accuracy (Agrawal et al., 2016). We evaluate VQA accuracy by comparing the agent's final predictions against ground-truth answers. We employ GPT-4o (OpenAI, 2024) as an automated judge to assign a binary score (1 for correct, 0 for incorrect) to each VQA pair.

We compute accuracy in both single-turn and multi-turn settings as described previously. Once all questions in the dataset have received a final prediction, either through early resolution or by reaching the maximum turn limit T. The overall VQA accuracy is calculated across the full test set across 3 independent runs, reporting 95% confidence intervals.

Component-wise evaluation We also report performance for separate components of CoA:

- **Controller performance:** We report accuracy, precision, recall, and F1 for whether the controller correctly predicts CLARIFY versus ANSWER over the mixed ambiguous and contrast sets.
- **Clarification quality:** Beyond VQA accuracy, we also evaluate clarification questions using the Ambiguity Resolution Reward metric introduced in our reward function $\gamma_{\text{resolution}}$ estimating whether answering the clarification question is likely to change the final VQA outcome. We additionally validate this reward metric with human judgments. We treat $\gamma_{\text{resolution}}$ as a diagnostic proxy for ambiguity resolution during policy training, relying on independent VQA accuracy as our primary metric to prevent reward hacking.

Human validation To ensure metric reliability, the first authors also manually annotate a sample of the test data across models for both VQA accuracy and the quality of generated clarification questions. We randomly sample and annotate 20% of the test dataset for VQA accuracy using the same binary score standard given to GPT-4o, and 30% of the CONTEXTCLARIFY test set to evaluate whether the generated clarification question resolves ambiguity from the original question. Human validation instructions and operationalization details can be found in Appendix F.

6 Results & Discussion

6.1 VQA Accuracy

Overall Performance Enhancement CoA improves overall VQA accuracy for the tested open-weight models over prompting and other baselines on CONTEXTCLARIFY (Table 1). Finetuned smaller models narrow the gap to frontier models (GPT-4o, Gemini-2.5-Flash),

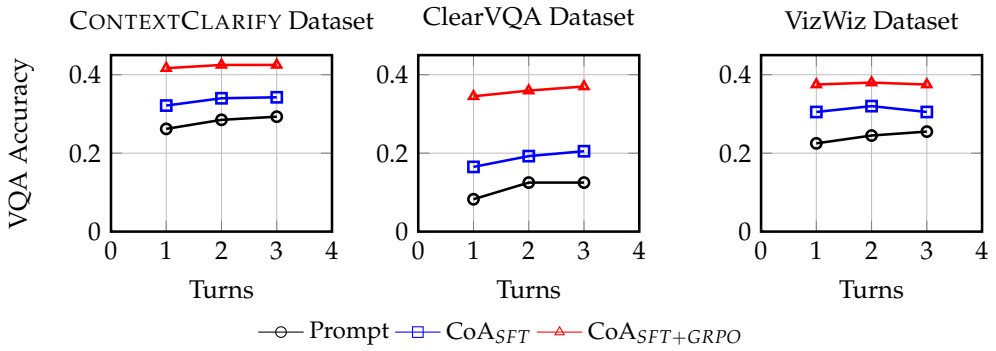


Figure 3: VQA accuracy comparison shows the most significant gains occurring within the first turn.

especially in categories such as *Cultural Background* or *Temporal Information*. While SFT provides consistent gains, GRPO-CR leads to larger improvements on average, outperforming zero-shot prompting, Always-Ask (AA$_{\text{GRPO}}$), and ICL baselines.

OOD Generalization Performance gains generalize to OOD datasets ClearVQA and VizWiz (Table 2). Across all three base models, CoA with GRPO yields consistent improvements to final VQA accuracy, even though both OOD datasets consist of different types of visual/textual ambiguity than those included in CONTEXTCLARIFY.

Effect of Multi-turn Clarification Figure 3 shows VQA accuracy as a function of the number of clarification turns. Most methods benefit from additional clarification turns, but the additional gains are marginal, and in a few settings actually lead to reduced overall accuracy. This suggests that the most critical ambiguities can be resolved early, and that extended interaction mostly leads to consistent or minor degradations in context.

Method	IVL-2B	Qw-3B	Qw-7B
ClearVQA			
Zero-shot	0.015	0.075	0.083
AA$_{\text{GRPO}}$	0.150	0.175	0.205
ICL	0.085	0.098	0.135
CoA$_{SFT}$	0.138	0.143	0.165
CoA$_{SFT+GRPO}$	**0.213**	**0.253**	**0.345**
CoA$_{\text{Oracle}}$	0.305	0.358	0.418
VizWiz			
Zero-shot	0.040	0.165	0.225
AA$_{\text{GRPO}}$	0.310	**0.380**	**0.395**
ICL	0.090	0.185	0.215
CoA$_{SFT}$	0.190	0.310	0.305
CoA$_{SFT+GRPO}$	**0.315**	0.360	0.375
CoA$_{\text{Oracle}}$	0.310	0.380	0.395

Table 2: OOD perf on ClearVQA and VizWiz. VizWiz contains only ambiguous VQAs, so AA is equivalent to Oracle.

Error Analysis We examine model decisions, clarification questions, and answers to understand how CoA$_{SFT+GRPO}$ behaves across ambiguous and unambiguous VQA scenarios compared to baselines. Figure 4 shows select examples (more in Appendix H). When questions are underspecified, CoA$_{SFT+GRPO}$ identifies missing contextual factors and generates targeted clarification questions that either fully resolve the ambiguity or narrow it to support a more reliable answer. In contrast, baseline prompting results in answers based on unwarranted inferences from partial visual cues. When questions are unambiguous, CoA$_{SFT+GRPO}$ avoids unnecessary clarification and demonstrates an ability to answer directly, indicating that gains generally stem from selective and effective clarification rather than always asking.

6.2 Module-level Performance

Ambiguity Detection and Controller Generalization Table 3 shows training the CoA controller with SFT consistently improves ambiguity detection over zero-shot prompting on CONTEXTCLARIFY and generalizes to the ClearVQA setting. The results show that the controller is much better at catching cases where a question is unclear, whereas the basic prompting method often misses them entirely. This is especially true for smaller models like IVL-2B, which struggle to trigger any clarifications without this specific training. Most importantly, these gains hold up even on the OOD ClearVQA data (we do not evaluate on VizWiz because it only contains ambiguous instances without a contrast set). This suggests that the model does not just memorize specific examples, but has learned general patterns for when to ask for more information.

Controller Bottleneck Analysis. To test how much controller errors cap end-to-end performance, we evaluate an *Oracle Controller* (CoA$_{\text{Oracle}}$) with ground-truth decision labels. As shown in Tables 1 and 2, replacing the learned controller with an oracle provides only

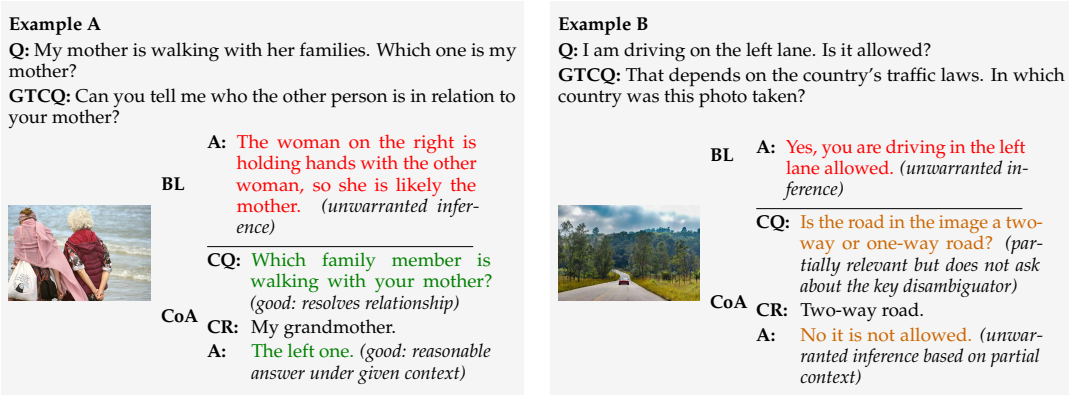


Figure 4: Select examples comparing responses from the prompting baseline (BL) and CoA$_{SFT+GRPO}$. In both examples, the BL answers without resolving the ambiguity. In Example A, CoA correctly resolves the ambiguity; in example B, while CoA produces a partially relevant clarifying question, it does not arrive at the correct answer. Abbreviations: Q = original question; GTCQ = ground truth clarification question; CQ = clarification question; CR = clarification response; A = final answer.

Data	Model	Method	Acc ↑	Prec	Recall ↑	F1 ↑
CtxClarify	IVL-2B	Prompt	0.500	0.000	0.000	0.000
		SFT	**0.536**	**0.588**	**0.238**	**0.339**
	Qw-3B	Prompt	0.476	0.474	0.429	0.450
		SFT	**0.500**	**0.500**	**0.500**	**0.500**
	Qw-7B	Prompt	0.595	**0.667**	0.381	0.485
		SFT	**0.702**	0.660	**0.833**	**0.737**
ClearVQA (OOD)	IVL-2B	Prompt	**0.495**	0.450	0.045	0.082
		SFT	0.490	**0.488**	**0.395**	**0.437**
	Qw-3B	Prompt	0.423	0.432	0.500	0.464
		SFT	**0.470**	**0.476**	**0.605**	**0.533**
	Qw-7B	Prompt	0.474	0.482	0.729	0.580
		SFT	**0.530**	**0.519**	**0.805**	**0.631**

Table 3: **Controller performance** on CONTEXTCLARIFY and ClearVQA. Prompt = zero-shot prompting. Training the CoA controller with SFT improves ambiguity detection on CONTEXTCLARIFY over Prompt and performance gains generalize to ClearVQA.

Dataset	Method	IVL-2B	Qw-3B	Qw-7B
CtxClarify	Prompt	0.476	0.518	0.571
	CoA$_{SFT}$	0.593	0.643	0.643
	CoA$_{SFT+GRPO}$	**0.685**	**0.768**	**0.810**
ClearVQA (OOD)	Prompt	0.526	0.561	0.660
	CoA$_{SFT}$	0.648	0.670	0.738
	CoA$_{SFT+GRPO}$	**0.653**	**0.796**	**0.804**
VizWiz (OOD)	Prompt	0.504	0.453	0.689
	CoA$_{SFT}$	0.611	0.559	0.708
	CoA$_{SFT+GRPO}$	**0.764**	**0.720**	**0.864**

Table 4: **Clarification question quality** across datasets and models. Scores range from 0 to 1. While SFT baseline provides a notable boost in precision over zero-shot prompting, our GRPO-CR method consistently achieves the highest clarification scores across both in and out domain datasets.

modest accuracy gains. This demonstrates that while ambiguity detection is non-trivial, the controller is not the main system bottleneck; downstream clarification quality and context integration play an equally critical role.

Clarification Question Quality Improvements CoA$_{SFT+GRPO}$ improves clarification question quality over both prompting and supervised fine-tuning, and generalizes to OOD datasets. Table 4 reports ambiguity resolution scores across all datasets and model families.

On CONTEXTCLARIFY, CoA$_{SFT+GRPO}$ achieves the best ambiguity resolution across all evaluated models, substantially improvingover both prompting and SFT. Gains are especially pronounced for Qwen-7B and InternVL-2B. These improvements also transfer to ClearVQA and Vizwiz, where CoA$_{SFT+GRPO}$ consistently yields the highest scores across models, showing that the learned clarification strategy does not overfit to CONTEXTCLARIFY and remains effective under distribution shift.

6.3 Human Validation

Human Evaluation of Generated Clarification Questions We conduct a human annotation (Appendix F) across three models on CONTEXTCLARIFY in Table 6. We found an 84% percentage agreement between human and GPT-based evaluations, corresponding to a Cohen's Kappa of 0.660. This indicates substantial agreement, confirming that the LLM judge is a reliable proxy for evaluating clarification quality.

Human Validation of LLM Judge To ensure reliability of our automated evaluation, we conduct a human correlation study (Appendix G) and calculate percent agreement (*PA*) between GPT-based scores and human judgments. Results demonstrate positive correlation ($PA \approx 82\%$). GPT evaluation is more conservative than human judgment. While humans use common sense to capture hidden context, the LLM is more literal and requires direct visual proof to mark an answer accurate. This shows that while our auto-eval metric introduces some error, it serves as a reasonable proxy for human perception on this judgment task.

7 Limitations & Future Work

Broader coverage and generalization. CONTEXTCLARIFY is derived from CODIS and focuses on single-factor underspecification. Extending coverage to more domains (e.g., diagrams, documents, embodied scenes), more diverse question styles, and more languages would improve external validity. Additionally, future datasets could explicitly annotate multiple plausible missing factors to study how clarification policies prioritize which question to ask first, and whether the agent can ask the most informative question under a one-question budget.

Method	IVL-2B	Qw-3B	Qw-7B
CtxClarify			
Zero-shot	0.105	0.236	0.316
$\text{CoA}_{SFT+GRPO}$	**0.368**	**0.316**	**0.474**
ClearVQA			
Zero-shot	0.276	0.233	0.263
$\text{CoA}_{SFT+GRPO}$	**0.333**	**0.341**	**0.375**

Table 5: Human evaluation of VQA accuracy on CtxClarify and ClearVQA.

Method	IVL-2B	Qw-3B	Qw-7B
Zero-shot	0.104	0.376	0.500
CoA_{SFT}	0.562	0.712	0.712
$\text{CoA}_{SFT+GRPO}$	**0.688**	**0.838**	**0.876**

Table 6: Human evaluation of clarification question quality. Questions generated by $\text{CoA}_{SFT+GRPO}$ on CtxClarify consistently achieve higher human evaluation scores across models.

End-to-end optimization and richer objectives. CoA currently optimizes the controller and clarification policy, while treating the answering model as a fixed backbone conditioned on the original inputs (or augmented with the clarification response). Joint optimization of all components could yield further gains but raises credit-assignment and stability challenges. Another direction is to optimize for user-centric objectives such as helpfulness, succinctness, and trust calibration, in addition to final answer accuracy.

Role of GPT-4o and real-world limitations. We rely heavily on GPT-4o to draft initial dataset candidates, compute similarity rewards during RL training, simulate user responses during testing, and judge VQA answer accuracy. To ensure quality and reduce bias, the authors manually review and refine all dataset questions, and validate a subset of responses for both RL rewards and VQA accuracy. While simulated responses allow for benchmarking feasibility, real human input is naturally noisier. Naturally, testing with human participants remains an important next step.

8 Conclusion

We present CoA (Clarify-or-Answer), an ask-or-answer agentic framework for context-dependent VQA that separates *when to ask* (controller) from *what to ask* (clarification policy), and produces a final answer either directly or after incorporating a single clarification response. We introduce CONTEXTCLARIFY, a dataset comprising 266 ambiguous instances with human-verified, open-ended clarification questions and 266 non-ambiguous contrast instances with complete context, enabling training and evaluation of ask-or-answer decisions and clarification generation. To learn effective clarification behavior, we propose GRPO-CR, a GRPO reinforcement learning method with multi-signal rewards encouraging focused, ambiguity-resolving questions. Our framework supports both single-turn and recursive multi-turn execution. The first turn provides the most performance gains and the recursive multi-turn paradigm further boosts accuracy by iteratively enriching the context. Across module and system level evaluations, and different model families, $\text{CoA}_{SFT+GRPO}$ achieves the highest overall VQA accuracy compared to zero-shot prompting, ICL, AA_{GRPO} and CoA_{SFT}, approaching the upper bound of an oracle controller ($\text{CoA}_{\text{Oracle}}$) and highlighting the value of explicit clarification for context-dependent visual and textual reasoning.

Acknowledgements

This work is partially supported by gift funds from Google and the Allen Institute for AI.

Ethical Statement

Potential Risks. This work studies clarification question generation for ambiguous VQA. While clarification can improve reliability, models may ask inappropriate or biased questions, increase user burden, or be misused to elicit sensitive information if deployed without safeguards. To mitigate these risks, our dataset excludes personally identifiable or sensitive content and focuses on benign contextual ambiguities. This work is intended for research use and should not be directly deployed in high stakes domains without additional safeguards and human oversight.

Use of AI Assistants. AI assistants were used to support paper writing, develop data-processing and evaluation scripts, simulate clarification responses, compute similarity-based rewards, and generate initial drafts of clarification questions that are subsequently reviewed and/or modified by the authors to ensure accuracy and clarity. All AI-generated content was reviewed and validated by the authors, who retain full responsibility for the dataset, methodology, and conclusions.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. *arXiv:1505.00468 [cs]*, 10 2016. URL https://arxiv.org/abs/1505.00468.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 07 2019. doi: 10.1145/3331184.3331265.

Chongyan Chen, Yu-Yun Tseng, Zhuoheng Li, Anush Venkatesh, and Danna Gurari. Acknowledging focus ambiguity in visual questions, 2025. URL https://arxiv.org/abs/2501.02201.

Gheorghe Comanici, Eric Bieber, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, 2025. URL https://arxiv.org/abs/2507.06261.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges, 2023. URL https://arxiv.org/abs/2311.03287.

Rachel N. Denison, William T. Adler, Marisa Carrasco, and Wei Ji Ma. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, 115:11090–11095, 10 2018. doi: 10.1073/pnas.1717720115.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images, 2025. URL https://arxiv.org/abs/2505.15879.

Ishmam Md Farhan, Ishmam Tashdeed, Talukder Asir Saadat, Md Hamjajul Ashmafee, A R M Kamal, and Md Azam Hossain. Visual robustness benchmark for visual question answering (vqa), 2024. URL https://arxiv.org/abs/2407.03386.

Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, Yunhang Shen, Mengdan Zhang, Peixian Chen, Sirui Zhao, Shaohui Lin, Deqiang Jiang, Di Yin, Peng Gao, Ke Li, Hongsheng Li, and Xing Sun. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv (Cornell University)*, 01 2023. doi: 10.48550/arxiv.2312.12436.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2023. URL https://arxiv.org/abs/2310.14566.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. Teaching vision-language models to ask: Resolving ambiguity in visual questions. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3619–3638, 2025. doi: 10.18653/v1/2025.acl-long.182. URL https://aclanthology.org/2025.acl-long.182/?utm_source=chatgpt.com.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Pedram Hosseini, Bryceton G Thomas, Jessica M Sin, Bing Ren, Jonathan S Ilgen, Yulia Tsvetkov, and Maarten Sap. Alfa: Aligning llms to ask good questions a case study in clinical reasoning, 2025. URL https://arxiv.org/abs/2502.14860.

Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, and Yang Liu. Codis: Benchmarking context-dependent visual comprehension for multimodal large language models. *arXiv preprint arXiv:2402.13607*, 2024.

OpenAI. Gpt-4o system card. 2024.

OpenAI, Aaron Hurst, Adam Lerer, et al. Gpt-4o system card. Technical report, 2024. URL https://arxiv.org/abs/2410.21276.

Ma Qian, S M Rayeed, Charles V Stewart, Qiong Wu, and Ma Yao. Identifying and resolving pitfalls of knowledge-based vqa benchmarks: Auditing, repairing, and augmenting, 2026. URL https://arxiv.org/abs/2607.00159.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 05 2023. URL https://arxiv.org/abs/2305.18290.

Sudha Rao and Hal Daumé. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *Meeting of the Association for Computational Linguistics*, 07 2018. doi: 10.18653/v1/p18-1255.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering, 2019. URL https://arxiv.org/abs/1902.05660.

Zhihong Shao, Runxin Xu, Peiyi Wang, Qihao Zhu, Junxiao Song, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and Yang Liu. Mucar: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models, 2025. URL https://arxiv.org/abs/2506.17046.

Y. Wang and K. Itou. Vqa: Detecting ambiguity and generating multiple answer candidates for clarifying visual questions. *Eighth International Conference on Sensors, Signal, and Image Processing (SSIP 2025)*, pp. 1, 3 2026. doi: 10.1117/12.3106042.

Bingbing Wen, Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Bill Howe, and Lijuan Wang. Infovisdial: An informative visual dialogue dataset by bridging large multimodal and language models. *arXiv preprint arXiv:2312.13503*, 2023.

Bingbing Wen, Bill Howe, and Lucy Lu Wang. Characterizing LLM abstention behavior in science QA with context perturbations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3437–3450, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.197. URL https://aclanthology.org/2024.findings-emnlp.197/.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Zhe Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=es2NkPKFCB.

Jihan Yao, Yushi Hu, Yujie Yi, Bin Han, Shangbin Feng, Guang Yang, Bingbing Wen, Ranjay Krishna, Lucy Lu Wang, Yulia Tsvetkov, et al. Mmmg: a comprehensive and reliable evaluation suite for multitask multimodal generation. *arXiv preprint arXiv:2505.17613*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Michael Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with lms, 2023. URL https://arxiv.org/abs/2311.09469.

Michael Zhang, Knox W Bradley, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions, 2024. URL https://arxiv.org/abs/2410.13788.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A GRPO Reward Function Parameter Sweep

We observe that clarification effectiveness remains relatively stable across moderate variations in reward scales, indicating that the proposed reward design is not highly sensitive to precise coefficient choices (Figure 11).

The picked configuration consistently performs near the top, supporting our empirical selection. Among the components, overly strong format rewards degrade performance, suggesting that enforcing structural correctness too aggressively can hinder informativeness, while weaker format rewards are sufficient as a soft constraint. Ambiguity and relevance rewards show stable behavior across settings, indicating that their contributions are important but not dependent on fine-grained tuning. Overall, these results suggest that the model benefits from a balanced combination of reward signals rather than precise optimization of individual coefficients.

Reward Scaling Rationale. We assign equal weights ($\lambda_k = 1.0$) across reward components to maintain a simple, un-tuned baseline and prevent overfitting to a specific reward configuration. Individual reward scales were bounded within a standard $[-1.0, +1.0]$ interval based on pilot experiments and established RL practice (Shao et al., 2024; Fan et al., 2025), ensuring that format compliance, relevance, and ambiguity resolution signals maintain comparable magnitudes during GRPO optimization.

B Details of Focused Relevance Reward

To address the ambiguity categories and the scoring criteria for the Focused Relevance Reward ($\gamma_{\text{relevance}}$), we provide the specific keyword mappings and linguistic patterns used during training.

Keywords Per Ambiguity Type A clarification question C_t satisfies the **Keywords** condition if it contains at least one token from the set associated with its specific ambiguity type. These categories and their respective keywords are defined in Table 7.

Category	Keywords
Cultural Background	country, city, state, region, location, place, area, nation, where, origin, cultural, tradition, custom, law, legal, illegal, rules, culture
Temporal Information	before, after, prior, following, earlier, later, recently, soon, now, current, time, date, day, night, morning, afternoon, evening, today, tomorrow, yesterday
Location & Orientation	left, right, up, down, forward, backward, toward, away, direction, push, pull, open, close, above, below, enter, exit
Attributes	hot, cold, on, off, high, low, big, small, large, short, long, heavy, light, temperature, speed, limit, safe, dangerous, bright, dark, clean, dirty, full, empty
Relationships	who, which, person, man, woman, boy, girl, friend, family, colleague, partner, teacher, student, owner, driver, passenger, speaker, listener, helper

Table 7: Mapping of Ambiguity Categories to Targeted Keywords

Clarification Patterns The pattern-matching condition ensures that questions are structurally oriented toward information seeking rather than generic conversational fillers.

- **Specific Patterns (Satisfies Condition)**: These identify questions targeting specific properties, alternatives, or comparative states:
  - Explicit Choice: e.g., *"either"*, *"neither"*, or the use of *"A or B"*.
  - Property Targeting: short wh-questions such as *"which [type/kind/part]?"* or *"what [version/value]?"*.
  - Reference Resolution: e.g., *"which of them?"* or *"what purpose?"*.
  - Comparative Cues: e.g., *"earlier"*, *"smaller"*, *"higher"*.
- **Vague Patterns (Fails Condition)**: These identify generic or non-specific requests that fail the structural relevance test:
  - Generic Requests: e.g., *"tell me more"*, *"elaborate"*, *"give me details"*.
  - Confusion Markers: e.g., *"I don't understand"*, *"not sure"*.
  - Indefinite Pronouns: e.g., *"anything"*, *"something"*.
  - Bare Why: a standalone *"why?"* without a following subject or object.

Reward Calculation The final score is assigned as follows:

$$\gamma_{\text{relevance}} = \begin{cases} +0.3 & \text{if both keyword and specific Pattern match} \\ +0.1 & \text{if only one condition matches} \\ -0.1 & \text{if neither condition matches} \end{cases} \tag{1}$$

Dataset	Model	Removed	Kept Rewards	Effectiveness
CtxClarify (Auto)	InternVL3-2B	None	GT + F + R + N + A	0.248 ± 0.010
CtxClarify (Auto)	InternVL3-2B	A	GT + F + R + N	0.145 ± 0.101
CtxClarify (Auto)	InternVL3-2B	N	GT + F + R + A	0.107 ± 0.006
CtxClarify (Auto)	InternVL3-2B	R	GT + F + N + A	0.126 ± 0.003
CtxClarify (Auto)	InternVL3-2B	A + N + R	GT + F	0.101 ± 0.008

Table 8: Ablation of GRPO-CR reward components. GT = Ground Truth similarity, F = Format, R = Relevance, N = Novelty, A = Ambiguity.

Split	# Samples	Percentage
Training	186	70%
Validation	40	15%
Test	40	15%
Total	266	100%

Table 9: Dataset split of CONTEXTCLARIFY for the ambiguous subset. The non-ambiguous subset uses the same split ratio and preserves the corresponding questions in each split.

C GRPO Reward Ablation

To evaluate the contribution of each reward component in GRPO-CR, we conduct an ablation study by selectively removing individual rewards while keeping the remaining components unchanged. We report clarification effectiveness on the CtxClarify dataset using InternVL3-2B.

Overall impact. As shown in Table 8, removing any individual reward component leads to a substantial drop in clarification effectiveness, confirming that the reward signals are complementary and jointly necessary for high quality clarification generation.

Ambiguity reward. Removing the ambiguity reward results in a large performance drop and significantly increased variance. This suggests that explicitly incentivizing ambiguity resolution is critical for stabilizing training and guiding the model toward asking clarifying questions rather than prematurely answering.

Novelty reward. The largest degradation is observed when removing the novelty reward, indicating that without encouraging diverse and non-trivial questions, the model tends to generate generic or redundant clarifications that fail to resolve uncertainty effectively.

Relevance reward. Removing the relevance reward also reduces performance, highlighting its role in ensuring that clarification questions remain grounded in the visual and textual context rather than drifting into unrelated or overly broad queries.

Minimal reward setting. When only Ground Truth similarity and Format rewards are retained, performance drops further, demonstrating that structural correctness alone is insufficient; semantic rewards (ambiguity, novelty, relevance) are essential for meaningful clarification.

Takeaway. These results validate our joint reward design in GRPO-CR. Each component addresses a distinct failure mode and their combination is necessary to achieve robust clarification behavior.

D Dataset Details

D.1 Data Split

Table 9 summarizes the data splits for the ambiguous subset of CONTEXTCLARIFY. The non-ambiguous subset follows the same split ratio.

(a) Q: My mother is walking with her families. Which one is my mother?
GTCQ: Can you tell me who the other person is in relation to your mother?

BL
A: The woman on the right is holding hands with the other woman, so she is likely the mother. *(unwarranted inference)*

CoA
CQ: Which family member is walking with your mother? *(good: resolves relationship)*
CR: My grandmother.
A: The left one. *(good: reasonable answer under given context)*

(b) Q: I am driving on the left lane. Is it allowed?
GTCQ: That depends on the country's traffic laws. In which country was this photo taken?

BL
A: Yes, you are driving in the left lane allowed. *(unwarranted inference)*

CoA
CQ: Is the road in the image a two-way or one-way road? *(partially relevant but does not ask about the key disambiguator)*
CR: Two-way road.
A: No it is not allowed. *(unwarranted inference based on partial context)*

(c) Q: Am I going from left to right or from right to left?
GTCQ: Is the door opening towards you or away from you?

BL
A: You are going from left to right. *(unwarranted inference)*

CoA
CQ: Is the door opening towards you or away from you? *(good: resolves temporal ambiguity)*
CR: Opening towards me.
A: From left to right. *(good: reasonable answer under given context)*

(d) Q: I want to buy two bottles of milk. Is my gesture appropriate?
GTCQ: In what country or region is this gesture being used?

BL
A: Yes, your gesture is appropriate *(unwarranted inference)*

CoA
CQ: What is the context or location where the gesture is being made? *(resolves the culture ambiguity)*
CR: In Greece.
A: No. *(good: resonable answer under the given context)*

(e) Q: Is the subway going in the same direction as I face?
GTCQ: Is the subway train moving towards or away from you?

BL
A: The image does not provide enough information to determine the direction the subway is going relative to the person facing the image. *(detects ambiguity but does not resolve it)*

CoA
CQ: Is the subway train moving towards or away from you? *(good: resolves orientation ambiguity)*
CR: Towards me.
A: No. *(good: reasonable answer under given context)*

(f) Q: The clock is on 12-hour format. What time is it on a 24-hour schedule?
GTCQ: Is this photo from the morning or late night?

BL
CQ: What is the current time on a 24-hour clock? *(repeats original question, no additional information gained)*
CR: It is 10:20 on a 24 hour schedule
A: 10:20.

CoA
CQ: Is the clock showing the time in the morning or afternoon? *(good:resolves AM/PM)*
CR: In the morning.
A: 10:20. *(good: reasonable answer under given context)*

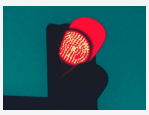
Figure 5: Case studies illustrating successful and unsuccessful ambiguity resolution. Each block compares the prompting baseline and CoA responses with annotations. Abbreviations: Q = original question; GTCQ = ground truth clarification question; CQ = clarification question; CR = clarification response; A = final answer.

(a) Q: My shop is going to be open. I am the owner of a shop. Should I flip the open sign or keep the position of the sign unchanged?
GT: Keep the position of the sign unchanged.

BL
A: Flip the open sign. *(wrong inference)*

CoA
A: Based on the image, it appears that the open sign is already flipped to indicate that the shop is open. Therefore, you should keep the position of the sign unchanged. *(reasonable inference)*

(b) Q: I am driving in New York. Can I make a right turn on the red traffic light?
GT: No.



BL
A: No, you cannot make a right turn on a red traffic light in new york. *(reasonable inference)*

CoA
A: No, you can not. The red light typically means "stop" and "do not proceed" in new york. *(reasonable inference)*

Figure 6: Case studies illustrating responses for un-ambiguous questions. Each block compares the prompting baseline and CoA with annotations. Abbreviations: Q = original question; GTCQ = ground truth clarification question; CQ = clarification question; CR = clarification response; A = final answer.

D.2 Ambiguity Types

Each ambiguous example in CONTEXTCLARIFY is associated with a single missing contextual factor. Following the categorization introduced in the original CODIS dataset (Luo et al., 2024), we consider five types of ambiguity:

- Location and orientation: the answer depends on geographic location or spatial orientation (e.g., country, hemisphere).
- Temporal information: the answer depends on time-related context (e.g., season, historical period).
- Cultural background: the answer depends on cultural or social norms.
- Attributes: the answer depends on properties of entities not visually specified (e.g., age, status).
- Relationships: the answer depends on social or interpersonal relationships.

Each ambiguous example is curated to isolate exactly one ambiguity type.

D.3 Ambiguity Category Breakdown and Annotation Setup

Table 10 provides a breakdown of CONTEXTCLARIFY across the five contextual ambiguity categories.

Attributes	Cultural Backgr.	Location & Orient.	Relationships	Temporal Info.	Total
116	86	114	84	132	532

Table 10: Distribution of instances across contextual ambiguity categories in CONTEXTCLARIFY.

Annotator Qualifications and Procedure. Dataset curation and manual verification were conducted by the paper's primary authors (graduate researchers in Computer Science and NLP). Candidate clarification questions generated during preprocessing were manually reviewed, edited, or re-annotated to ensure precise visual grounding and single-factor targeting. Disagreements were resolved through consensus discussion among the authors.

Balanced Dataset Design. We intentionally construct a balanced $1:1$ mixture of ambiguous and context-complete instances in our contrast set. This setup provides a clean baseline to evaluate the controller's ask-or-answer decision boundary without inducing degenerate strategies (e.g., always asking or never asking). While real-world deployments may present skewed ambiguity distributions, a balanced split provides a rigorous benchmark for selective clarification.

D.4 Mixed Dataset Construction Rules

CONTEXTCLARIFY is constructed deterministically from an initial set of ambiguous VQA items. For each ambiguous item, we construct one additional non-ambiguous variant by modifying the question text:

Context Completion We add the missing relevant context explicitly to the original question, yielding a fully specified VQA instance. These examples are labeled as `no_clarification_needed`, since the answer can be determined directly from the image and provided context.

Irrelevant Context Injection We add additional contextual information that is syntactically plausible but does not resolve the underlying ambiguity. These examples preserve the original ambiguity and are labeled as `needs_clarification`. Irrelevant context is selected to avoid introducing new ambiguities or revealing the correct answer.

This augmentation procedure results in a balanced mixture of ambiguous and non-ambiguous questions while preserving the original visual content.

D.5 Examples

Example 1 (Location Ambiguity)

Original question	Is this behavior legal?
Context-completed	Is this behavior legal in Germany?
Irrelevant context	Is this behavior legal while wearing a blue jacket?

Example 2 (Temporal Ambiguity)

Original question	Can this vehicle be parked here?
Context-completed	Can this vehicle be parked here on Sundays?
Irrelevant context	Can this vehicle be parked here during the afternoon?

E Training Details

E.1 GRPO-CR Training

We implemented GRPO-CR on top of the open-sourced GRPO training framework from GRIT (Fan et al., 2025). All GRPO experiments are conducted on 4×A100 (80GB) GPUs using distributed training with DeepSpeed ZeRO-2 or ZeRO-3.

We train for a single epoch over the CONTEXTCLARIFY ambiguous set training split. Each training instance generates multiple clarification candidates, which are optimized using preference-based rewards described in Section 3.4. We restrict the interaction to a single clarification turn.

Optimization. We use the AdamW optimizer with a learning rate of 2×10^{-6} and a cosine learning rate scheduler. The per-device training batch size is 2 with gradient accumulation over 2 steps. We set the KL regularization coefficient to $\beta = 0.01$. Training is performed in bfloat16 precision.

Generation and context limits. For each prompt, we sample 2 clarification generations. The maximum prompt length is set to 1000 tokens and the maximum completion length to 768 tokens.

Evaluation and checkpointing. Evaluation is performed every 50 steps, and checkpoints are saved every 20 steps, keeping up to 5 checkpoints. All training statistics and generated clarifications are logged to Weights & Biases.

E.2 Supervised Fine-Tuning

For supervised baselines, we fine-tune the same backbone model using the MS-SWIFT framework (Zhao et al., 2024). We adopt parameter-efficient fine-tuning with LoRA and train on the CONTEXTCLARIFY.

Setup. SFT is conducted on 2*A100(80GB) GPUs with DeepSpeed ZeRO-2. We use a LoRA rank of 8 and $\alpha = 32$, applying LoRA to all linear layers. The model is trained for 3 epochs with a learning rate of 1×10^{-4} and a warmup ratio of 0.05.

Batching and precision. The per-device batch size is 1 with gradient accumulation to achieve an effective batch size of 16. All SFT experiments are trained in bfloat16 precision, with a maximum sequence length of 4096 tokens.

Evaluation and checkpointing. We evaluate and save checkpoints every 100 steps, keeping the two most recent checkpoints.

E.3 In-Context Learning (ICL) Baseline Setup

To evaluate whether improved prompting enhances zero-shot behavior, we construct a 4-shot ICL baseline. Each prompt augments the zero-shot instruction with:

Reward Component	Setting	Clarification Effectiveness
Ambiguity Reward Sweep		
Ambiguity	$+0.7, -0.5$	0.925 ± 0.118
Ambiguity	$+0.5, -0.3$ (chosen)	0.938 ± 0.090
Ambiguity	$+0.3, -0.1$	0.938 ± 0.039
Novelty Reward Sweep		
Novelty	$+1.0, -0.1, -1.0$	0.730 ± 0.093
Novelty	$+0.3, -0.1, -0.3$ (chosen)	0.852 ± 0.108
Novelty	$+0.1, +0.0, -0.1$	0.820 ± 0.088
Format Reward Sweep		
Format	$+1.0, -1.0$	0.710 ± 0.205
Format	$+0.5, -0.5$ (chosen)	0.750 ± 0.090
Format	$+0.1, -0.1$	0.805 ± 0.205
Relevance Reward Sweep		
Relevance	$+0.5,\ +0.1,\ -0.3$	0.831 ± 0.130
Relevance	$+0.3,\ +0.1,\ -0.1$ (chosen)	0.831 ± 0.030
Relevance	$+0.1,\ +0.0,\ -0.1$	0.796 ± 0.090

Table 11: Parameter sweep over GRPO-CR reward coefficients. Each sweep varies one reward component while keeping all others fixed. Clarification effectiveness is computed using the ambiguity resolution reward and normalized to $[0, 1]$ for comparability across settings. Results are reported on CtxClarify (Auto) using InternVL3-2B.

- 1 context-sufficient example requiring a direct answer.
- 3 context-missing examples randomly sampled from the training set, covering *Cultural Background*, *Location & Orientation*, and *Temporal Information* ambiguity categories.

Each few-shot example provides the image, question, and ground-truth output (either direct answer or target clarification question). While ICL slightly improves over zero-shot prompting, it remains behind parameter-tuned models (CoA_{SFT} and $\text{CoA}_{\text{SFT+GRPO}}$).

F Human Validation of Ambiguity Resolution

To validate that the proposed automatic ambiguity resolution reward reflects human judgment, we conduct a manual evaluation on a subset of the CONTEXTCLARIFY test set. Human annotations were performed by the authors.

F.1 Sampling Protocol

We randomly sample 30% of the CONTEXTCLARIFY test set. For each sampled example, we evaluate clarification questions generated by three methods:vanilla prompting, supervised fine-tuning, and GRPO-CR. The same sampled examples are used across all methods and model variants to ensure a fair comparison.

F.2 Annotation Task

Human annotators are presented with the following information for each example:

- the image,
- the original ambiguous question,
- the model generated clarification question,
- the ground truth annotated clarification question, answer and final answer.

Annotators are asked to assess the extent to which the clarification question, when answered, resolves the ambiguity in the original question. Each clarification is assigned a graded score:

- **0 (Not Resolved):** The clarification does not address the source of ambiguity or would not change the final answer.
- **0.5 (Partially Resolved):** The clarification addresses part of the ambiguity but leaves multiple plausible answers.

- **1 (Fully Resolved):** The clarification directly targets the missing information and would uniquely determine the final answer.

G VQA Final Answer Accuracy Human Evaluation

We conduct a human evaluation to assess the correctness of the final VQA answers produced by different methods, focusing on whether the final response is reasonable for the given VQA. Human annotations were performed by the authors.

G.1 Sampling Protocol

We randomly sample 20% of the test set predictions for human annotation. Sampling is performed uniformly across methods to ensure fair and unbiased comparison.

G.2 Annotation Task

For each evaluated example, annotators are provided with:

- the image,
- the original question,
- the model's final answer,
- the clarification process, including the clarification question and response, when applicable

Each example is assigned a binary score: **1** if the final answer is reasonable, and **0** otherwise.

G.3 LLM Judge Discrepancy Analysis

Our human validation study indicates an 82% agreement rate with the GPT-4o judge. Qualitative error inspection reveals that disagreements stem primarily from the LLM judge applying a stricter, more conservative evaluation standard than human annotators:

1. **Strict Formatting Constraints:** The LLM judge penalizes minor formatting variations. For example, answering "1010" for a 24-hour clock question was rejected by the LLM for missing colons, whereas human annotators accepted it as semantically equivalent to "10:10".
2. **Overly Strict Evidence Requirements:** The LLM judge frequently penalizes responses relying on lightweight commonsense or conversational grounding. In a workplace query ("I am making a report to my leader"), the answer "the person on the left is the leader" was accepted by humans as reasonable contextual grounding, but rejected by the LLM for lacking explicit visual proof.
3. **Conservative Domain Interpretation:** In financial queries where green text implied stock gains, human annotators accepted "the stock is performing well", while the LLM judge rejected it due to demanding additional external market context.

H Additional Error Analysis

Figure 5 includes examples of ambiguous questions in our dataset, with the question and the ground-truth clarifying question, along with model responses for the prompting baseline and CoA trained with GRPO-CR. In the vast majority of cases, the prompting baseline answers directly (incorrectly) without seeking additional clarification. The baseline also occasionally fails by reproducing a variant of the original question. For CoA, the model asks a relevant clarification question, which resolves the ambiguity in some cases. Figure 6 shows case studies for non-ambiguous questions from our contrast set.

I Prompt Templates

We reproduce the prompts used for ambiguity detection, clarification question generation, and the end-to-end prompt baseline below.

I.1 Ambiguity Detection

```
System:
You are a visual Q&A ambiguity checker. Given one image and one question
```

```
about that image, decide if the question needs further clarification
before it can be reliably answered from pixels. Mark `yes` if answering
depends on off-image knowledge. Otherwise, mark `no`.

Output only a single token: `yes` or `no`.

User:
<Image>
Question: {question}
```

I.2 Clarification Question Generation

```
System:
You write exactly one clarifying question. Do not answer the original
question.

Given an image and an original question, ask for the single missing fact
that would determine the answer.

Input: (1) image, (2) original question.
Output: exactly one short clarification question without any prefix.

User:
<Image>
Question: {question}
```

I.3 Final Answer Generation

```
System:
You are a visual assistant. Answer the user's question using the image.
Input: (1) image, (2) original question.
Output: final answer

User:
<Image>
Question: {question}
```

I.4 Vanilla Prompt

```
System:
You are a visual assistant. You answer the question directly whenever
the image provides enough information. Only ask a clarification question
if the answer would differ depending on missing information.

Input: (1) image, (2) original question.
Output: final answer or a single short clarification question

User:
<Image>
Question: {question}
```

I.5 VQA Accuracy LLM Judge

```
System:
You are a strict but fair VQA (Visual Question Answering) judge.

Your job: decide whether the provided answer(s) correctly answer the
given question and image.

Rules:
- Output ONLY valid JSON, no preamble, no markdown fences.
- Output format: {"verdict": "correct" | "incorrect",
"reasoning": "<one sentence>"}
- "correct"   → the answer is a valid, reasonable response to the
question and image pair.
```

```
- "incorrect" → the answer is wrong, irrelevant, or does not address the
question and image pair.
- Be strict about factual accuracy; do not give credit for vague
non-answers.

User:
<Image>
Question: {question}
Answer: {answer}
```